Remark 2.5. In a Bernoulli trial, the number treated $N_1 = \sum_i A_i \sim \text{Bin}(n, \pi)$ is random, not fixed. In this section we also view the potential outcomes as random, not fixed.

Recall the difference-in-means estimator is given by

$$\widehat{\psi} = \frac{1}{\sum_i A_i} \sum_{i:A_i=1} Y_i - \frac{1}{\sum_i (1-A_i)} \sum_{i:A_i=0} Y_i = \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)}$$

Now we will study the properties of $\widehat{\psi}$: bias, variance, and limiting distribution. We will see that very precise estimation and inference are possible for the causal effect ψ in Bernoulli trials, under essentially no assumptions beyond consistency.

2.4.1 Properties of the Difference-in-Means Estimator

Theorem 2.1. *Assume consistency. In a Bernoulli trial, the difference-in-means estimator is unbiased for ψ and has variance no greater than*

$$\frac{2}{(n+1)} \left(\frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi} \right)$$

where $\sigma_a^2 = var(Y \mid A = a)$.

Proof. Let $\widehat{\pi} = \mathbb{P}_n(A)$ and just consider the first term $\widehat{\mu}_1 = \mathbb{P}_n(AY)/\widehat{\pi}$ as an estimator of $\mu_1 = \mathbb{E}(Y \mid A = 1)$. We have

$$\begin{aligned} \mathbb{E}(\widehat{\mu}_1 \mid A^n) &= \frac{1}{\widehat{\pi}} \mathbb{E}\Big\{ \mathbb{P}_n(AY) \mid A^n \Big\} = \frac{1}{\widehat{\pi}} \mathbb{P}_n \Big\{ A\mathbb{E}(Y \mid A^n) \Big\} \\ &= \frac{1}{\widehat{\pi}} \mathbb{P}_n \Big\{ A\mathbb{E}(Y \mid A = 1) \Big\} = (\widehat{\pi}\mu_1)/\widehat{\pi} = \mu_1 \end{aligned}$$

where the third equality used the iid assumption. (Note: why did we not have to use randomization here? Or where did we implicitly use it?). Unbiasedness now follows by iterated expectation, and consistency follows from the weak law of large numbers and continuous mapping theorem. The logic is the same for $\widehat{\mu}_0 = \mathbb{P}_n\{(1-A)Y\}/(1-\widehat{\pi})$.

By the law of total variance we have

$$\text{var}(\widehat{\mu}_1) = \text{var}\Big\{ \mathbb{E}(\widehat{\mu}_1 \mid A^n) \Big\} + \mathbb{E}\Big\{ \text{var}(\widehat{\mu}_1 \mid A^n) \Big\}$$

Note $\text{var}\{\mathbb{E}(\widehat{\mu}_1 \mid A^n)\} = \text{var}(\mu_1) = 0$ from above, and

$$\begin{aligned} \text{var}(\widehat{\mu}_1 \mid A^n) &= \left(\frac{1}{n\widehat{\pi}} \right)^2 \sum_{i=1}^n A_i \text{var}\left(Y_i \mid A^n \right) \\ &= \left(\frac{1}{n\widehat{\pi}} \right)^2 \sum_{i=1}^n A_i \sigma_1^2 = \frac{\sigma_1^2}{N_1} \mathbb{1}(N_1 > 0) \end{aligned}$$

where we used independence and defined $\sigma_1^2 = \text{var}(Y \mid A = 1)$ and $N_1 = n\widehat{\pi}$. Now

$$\begin{aligned} \text{var}(\widehat{\mu}_1) &= \mathbb{E}\Big\{\text{var}(\widehat{\mu}_1 \mid A^n)\Big\} \\ &\leq \frac{2\sigma_1^2}{(n+1)\pi} \end{aligned}$$

by the expected binomial reciprocal result (Lemma A.2) of Devroye et al. [1996]. The same logic applies to $\widehat{\mu}_0$, and iterated expectation shows that the covariance term $\text{cov}(\widehat{\mu}_1, \widehat{\mu}_0)$ is exactly zero, which gives the result. □

Theorem 2.2. *Assume consistency. For a Bernoulli trial, the difference-in-means estimator is root-n consistent and asymptotically normal with*

$$\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N\left(0, \ \frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi}\right)$$

where $\sigma_a^2 = var(Y \mid A = a)$.

Proof. We again focus on μ_1 and its estimator. Note we have

$$\begin{aligned} \widehat{\mu}_1 - \mu_1 &= \frac{\mathbb{P}_n(AY)}{\widehat{\pi}} - \mu_1 = \mathbb{P}_n\left\{\frac{A}{\widehat{\pi}}(Y - \mu_1)\right\} \\ &= \mathbb{P}_n\left\{\frac{A}{\pi}(Y - \mu_1)\right\} + \left(\frac{1}{\widehat{\pi}} - \frac{1}{\pi}\right)\mathbb{P}_n\left\{A(Y - \mu_1)\right\} \\ &= \mathbb{P}_n\left\{\frac{A}{\pi}(Y - \mu_1)\right\} + O_{\mathbb{P}}(1/\sqrt{n})O_{\mathbb{P}}(1/\sqrt{n}) \end{aligned}$$

where the last equality follows by the central limit theorem, which implies $\sqrt{n}\{\mathbb{P}_n(V) - \mathbb{E}(V)\} = O_{\mathbb{P}}(1)$ for any iid V with finite mean and variance, together with the fact that $(\widehat{\pi}, \pi)$ are bounded away from zero.

Therefore

$$\widehat{\mu}_1 - \mu_1 = \mathbb{P}_n\left\{\frac{A}{\pi}(Y - \mu_1)\right\} + o_{\mathbb{P}}(1/\sqrt{n})$$

since $O_{\mathbb{P}}(1/\sqrt{n})O_{\mathbb{P}}(1/\sqrt{n}) = O_{\mathbb{P}}(1/n) = o_{\mathbb{P}}(1/\sqrt{n})$, which from the central limit theorem (together with Slutsky's theorem) gives

$$\sqrt{n}\Big(\widehat{\mu}_1 - \mu_1\Big) \rightsquigarrow N\left(0, \text{var}\left\{\frac{A}{\pi}(Y - \mu_1)\right\}\right)$$

The logic for the $\widehat{\mu}_0$ part is analogous.

□

Theorem 2.1 is powerful in showing that, in Bernoulli trials, mean counterfactuals can be estimated very precisely (i.e., with zero bias and variance that scales like $1/n$) using no assumptions other than consistency and finite variance. In other words: randomization allows accurate and essentially assumption-free causal inference!

Similarly, Theorems 2.1 and 2.2 also pave the way for inference, in the form of confidence intervals and hypothesis tests. Namely, finite sample confidence intervals could be constructed based on Theorem 2.1 using bounds on the conditional variances σ_a^2, and Theorem 2.2 implies for example that an asymptotic 95% CI is given by

$$\widehat{\psi} \pm \left(\frac{1.96}{\sqrt{n}}\right) \widehat{\mathrm{sd}} \left\{ \frac{A(Y-\widehat{\mu}_1)}{\pi} - \frac{(1-A)(Y-\widehat{\mu}_1)}{1-\pi} \right\}.$$

Remark 2.6. We saw above that the asymptotic variance of the difference-in-means estimator in a Bernoulli experiment is given by

$$\frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi}.$$

One interesting thing to note about this variance comes the perspective of experimental design: what is the best choice of π for optimizing efficiency? In fact, it is easy to show that

$$\arg\min_{\pi} \left(\frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi} \right) = \frac{\sigma_1}{\sigma_0 + \sigma_1}$$

so for optimal efficiency the proportion treated should match the standard deviation of treated outcomes, as a fraction of the total standard deviation for treated and untreated outcomes. This is intuitive – if outcomes are more variable among treated patients than controls (i.e., $\sigma_1 > \sigma_0$) then more patients should be assigned to treatment to counterbalance the extra noise.

2.4.2 Sample versus Population Effects

Here we point out an interesting connection between sample effect estimation in completely randomized experiments and population effect estimation in Bernoulli experiments.

Based on Theorem 2.2, an asymptotic 95% CI for ψ in a Bernoulli experiment is given by

$$\widehat{\psi} \pm 1.96 \sqrt{\frac{\widehat{\sigma}_1^2}{n\widehat{\pi}} + \frac{\widehat{\sigma}_0^2}{n(1-\widehat{\pi})}}$$

where $\widehat{\sigma}_a^2 \equiv \sigma_n^2(y^a)$ is the usual sample variance among the treated ($a = 1$) and controls ($a = 0$), which we used in our analysis of the difference-in-means as an estimator of the *sample* average effect in completely randomized experiments (e.g., Proposition 2.3).

In fact, $\widehat{\psi}$ is the exact same point estimate of the sample effect that we analyzed in completely randomized experiments, and similarly the exact same confidence interval

$$\widehat{\psi} \pm 1.96 \sqrt{\frac{\widehat{\sigma}_1^2}{n\widehat{\pi}} + \frac{\widehat{\sigma}_0^2}{n(1-\widehat{\pi})}}$$

is also valid (possibly conservative) in completely randomized experiments, guaranteeing at least 95% coverage of the sample effect. (This results from using the naive bound of $\sigma_n^2(y^1 - y^0) \geq 0$ as in (2.2)).

Thus, not only is the estimator for the population effect exactly the same as that for the sample effect, but confidence intervals for the population effect are also valid for the sample effect, being at worst conservative. This is an archetypal example of how finite-sample and population-based frameworks can coincide.

Note that, although population-based confidence intervals are valid for sample effects, the converse is not necessarily true: it is easier to estimate sample effects, in the sense that the same estimators have smaller variances relative to sample versus population effects. Thus a confidence interval for a sample effect may not be valid for a population effect. For example, Imbens [2004] shows that

$$\mathbb{E}\{(\widehat{\psi} - \psi_n)^2\} = \mathbb{E}\{(\widehat{\psi} - \psi)^2\} - \frac{\text{var}(Y^1 - Y^0)}{n} + o(1/n)$$

so that the difference-in-means has smaller variance when estimating the sample effect ψ_n. For some intuition, imagine both potential outcomes were observed for each subject: then the sample effect would be estimated without error, but not the population effect.

2.4.3 Difference-in-Means versus Horvitz-Thompson

Note that the difference-in-means estimator is given by

$$\widehat{\psi} = \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)} = \mathbb{P}_n\left\{\left(\frac{A}{\widehat{\pi}}\right)Y - \left(\frac{1-A}{1-\widehat{\pi}}\right)Y\right\}$$

which suggests a different version, where we replace the estimated proportion treated $\widehat{\pi}$ with its known population value π:

$$\widehat{\psi}_{ht} = \mathbb{P}_n\left\{\left(\frac{A}{\pi}\right)Y - \left(\frac{1-A}{1-\pi}\right)Y\right\}$$

This estimator is known as the Horvitz-Thompson estimator, hence the *ht* subscript.

Since we are replacing an estimated quantity $\widehat{\pi}$ with its known value π, it seems as if we should gain efficiency. Is this actually true?

It is straightforward to check that the Horvitz-Thompson estimator is also unbiased and consistent; and since it is exactly equal to a sample average, we can apply the central limit theorem to immediately obtain

$$\sqrt{n}(\widehat{\psi}_{ht} - \psi) \rightsquigarrow N\left(0, \text{var}\left\{\left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)Y\right\}\right)$$

Now which estimator should we use: difference-in-means or Horvitz-Thompson? Both are unbiased, root-n consistent, and asymptotically normal. Is our intuition correct that it is beneficial to replace the estimate $\widehat{\pi}$ with its known value π? To answer this we will compare asymptotic variances.

Let $\phi = \frac{A}{\pi}(Y - \mu_1) - \frac{1-A}{1-\pi}(Y - \mu_0)$ and $\phi_{ht} = \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right) Y$ denote the functions whose variances correspond to the asymptotic variances of $\widehat{\psi}$ and $\widehat{\psi}_{ht}$. Then we have

$$\begin{aligned} \text{var}(\phi_{ht}) &= \text{var}\left(\phi + \frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right) \\ &= \text{var}(\phi) + \text{var}\left(\frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right) \end{aligned}$$

where the last line follows since $\mathbb{E}(\phi \mid A) = 0$ implies that

$$\text{cov}\left(\phi, \frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right) = 0$$

by iterated expectation.

Therefore

$$\text{var}(\phi_{ht}) \geq \text{var}(\phi)$$

and thus the Horvitz-Thompson estimator is *less efficient* than the difference-in-means. This is counterintuitive: here replacing an estimated quantity with its known population counterpart actually reduces efficiency! Usually when we estimate things we get something *less precise* than if we just used the true quantity.

Unfortunately, I do not know of a very satisfying intuitive explanation of this paradox. One way to think about it is as follows: rather than viewing $\widehat{\psi}_{ht}$ as replacing an estimated quantity with a known quantity, one can instead view it as moving away from the sample average $\widehat{\psi} = \widehat{\mu}_1 - \widehat{\mu}_0$ with a noisier version

$$\widehat{\psi}_{ht} = \left(\frac{\widehat{\pi}}{\pi}\right)\widehat{\mu}_1 - \left(\frac{1-\widehat{\pi}}{1-\pi}\right)\widehat{\mu}_0$$

which should degrade performance, merely since sample averages are efficient estimators of means. In other words, the Horvitz-Thompson estimator is using the expected number of treated $n\pi$ rather than the actual number $n\widehat{\pi}$, so that when the actual number differs from its expectation, the averages are not correctly weighted.

Chapter 3

Randomized Experiments with Covariates

3.1 Identification with Covariates

So far we have considered settings where we have access to an iid sample

$$(A_1, Y_1), ..., (A_n, Y_n) \sim \mathbb{P}$$

but it is very common to also observe auxiliary covariate information (e.g., demographics like age or gender, or baseline outcome measures, etc.). Thus in practice we often have an iid sample

$$(X_1, A_1, Y_1), ..., (X_n, A_n, Y_n) \sim \mathbb{P}$$

for covariates or features $X \in \mathbb{R}^d$.

For now we will continue to pursue the experimental setting in which we can assume

1. consistency: $Y = AY^1 + (1 - A)Y^0$
2. randomization: $A \perp\!\!\!\perp (X, Y^a)$ for $a \in \{0, 1\}$ with $\mathbb{P}(A = 1 \mid X) = \pi$

Our goal is still to estimate the average treatment effect

$$\psi = \mathbb{E}(Y^1 - Y^0),$$

i.e., the mean outcome in the population if all versus none were treated. The questions we consider here are: Does our identification result $\psi = \mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0)$ without covariates still hold? Are there any new identification results that the covariates buy us? We will see that the answer to both questions is: yes.

We will require the following standard independence result.

Proposition 3.1. *If $U \perp\!\!\!\perp (V, W)$ then $U \perp\!\!\!\perp W$ and $U \perp\!\!\!\perp V \mid W$.*

Proof. Since $U \perp\!\!\!\perp (V, W)$ means $p(u, v, w) = p(u)p(v, w)$ we have

$$p(u, w) = \int p(u, v, w) \ dv = \int p(u)p(v, w) \ dv = p(u)p(w)$$

so that $U \perp\!\!\!\perp W$, showing the first part. For the second part note

$$p(u, v \mid w) = p(u \mid v, w)p(v \mid w) = p(u)p(v \mid w) = p(u \mid w)p(v \mid w).$$

□

Proposition 3.2. *Assume consistency and randomization as given above. Then*

$$\begin{aligned}\mathbb{E}(Y^1 - Y^0) &= \mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0) \\ &= \mathbb{E}\Big\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\Big\}\end{aligned}$$

where

$$\mathbb{E}\Big\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\Big\} = \int \Big\{\mathbb{E}(Y \mid X = x, A = 1) - \mathbb{E}(Y \mid X = x, A = 0)\Big\} \ d\mathbb{P}(x).$$

Proof. Proposition 3.1 tells us that $A \perp\!\!\!\perp (X, Y^a) \implies A \perp\!\!\!\perp Y^a$ and $A \perp\!\!\!\perp Y^a \mid X$, and we already know $A \perp\!\!\!\perp Y^a$ implies

$$\mathbb{E}(Y^a) = \mathbb{E}(Y^a \mid A = a) = \mathbb{E}(Y \mid A = a)$$

by randomization and consistency. For the second identification result note

$$\mathbb{E}(Y^a) = \mathbb{E}\{\mathbb{E}(Y^a \mid X)\} = \mathbb{E}\{\mathbb{E}(Y^a \mid X, A = a)\} = \mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\}$$

where the first equality follows by iterated expectation, the second by $A \perp\!\!\!\perp Y^a \mid X$, and the third by consistency. □

The two identification results above suggest (at least) two different estimators for, for example, $\psi_1 = \mathbb{E}(Y^1)$, namely:

$$\widehat{\psi}_1 = \mathbb{P}_n(Y \mid A = 1) \quad \text{versus} \quad \widehat{\psi}_1 = \mathbb{P}_n\{\widehat{\mathbb{E}}(Y \mid X, A = 1)\}$$

The questions we pursue here are: Which estimator is "better"? Should we incorporate the covariate information? How?

3.2 Logistic Regression & Collapsibility

For the time being, suppose $Y \in \{0, 1\}$ is a binary outcome. Maybe the most common approach in practice is to assume the logistic regression model

$$\text{logit } \mathbb{P}(Y = 1 \mid X, A) = \beta_0 + \beta_1 A + \beta_2^{\mathrm{T}} X$$

and call β_1 "the effect" of treatment. What "effect" does this actually represent?

First: this is likely not an effect at all, because *the model is probably wrong*. For better or worse, nature does not care about our logistic regressions. We fit logistic regression models because they are fast and easy, not because they are realistic. In reality any given model probably leaves out important covariate interactions, higher-order terms, covariate-treatment interactions, non-logit links, etc.

In other words, it is pretty presumptuous to assume we know the *exact* functional form explaining how covariates relate to the outcome, up to some finite-dimensional parameter. Nature is probably too complex for that. And thus if the model is wrong, β_1 is hard to interpret and not so meaningful – a projection at best.

Nevertheless, for the sake of argument assume that the logistic model is correct. Then

$$\exp(\beta_1) = \frac{\text{odds}(Y = 1 \mid X, A = 1)}{\text{odds}(Y = 1 \mid X, A = 0)} = \frac{\text{odds}(Y^1 = 1 \mid X)}{\text{odds}(Y^0 = 1 \mid X)}$$

where in the second equality we used consistency and randomization.

This is a *conditional odds ratio* (OR). Importantly

$$\mathbb{E}(Y^1 - Y^0) = \mathbb{P}(Y^1 = 1) - \mathbb{P}(Y^0 = 1) \neq \frac{\text{odds}(Y^1 = 1 \mid X)}{\text{odds}(Y^0 = 1 \mid X)}$$

so it is certainly not an average treatment effect (risk difference). In fact, *even if the model is correct*

$$\frac{\text{odds}(Y^1 = 1)}{\text{odds}(Y^0 = 1)} \neq \frac{\text{odds}(Y^1 = 1 \mid X)}{\text{odds}(Y^0 = 1 \mid X)}$$

so it is not even a population odds ratio effect (even if the conditional OR is constant!). This follows since

$$\begin{aligned}
\text{odds}(Y^1 = 1) &= \frac{\mathbb{P}(Y^1 = 1)}{\mathbb{P}(Y^1 = 0)} = \frac{\mathbb{P}(Y = 1 \mid A = 1)}{\mathbb{P}(Y = 0 \mid A = 1)} \\
&= \frac{\mathbb{E}\{\mathbb{P}(Y = 1 \mid X, A = 1)\}}{\mathbb{E}\{\mathbb{P}(Y = 0 \mid X, A = 1)\}} = \frac{\mathbb{E}\{\text{expit}(\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} X)\}}{1 - \mathbb{E}\{\text{expit}(\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} X)\}} \\
&\neq \frac{\text{expit}\{\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} \mathbb{E}(X)\}}{1 - \text{expit}\{\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} \mathbb{E}(X)\}} = \exp\{\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} \mathbb{E}(X)\}
\end{aligned}$$

since $\mathbb{E}\{f(X)\} \neq f\{\mathbb{E}(X)\}$ for nonlinear f. This is called the problem of *non-collapsibility* [Freedman, 2008, Greenland et al., 1999]. We say the odds ratio is not collapsible since the average of the conditional ORs is not generally equal to the marginal OR.

In fact the marginal OR can be bigger/small than all of the conditional ORs. This is counterintuitive! Consider an example with half men and half women:

- men have 20% chance of heart attack when treated, 50% when not
- women have 2% chance of heart attack when treated, 8% when not

What are the ORs in this setup?

$$\begin{aligned}
OR(HA \mid men) &= \frac{\text{odds}(HA^1 \mid men)}{\text{odds}(HA^0 \mid men)} = \frac{0.2/0.8}{0.5/0.5} = 0.25 \\
OR(HA \mid women) &= \frac{\text{odds}(HA^1 \mid women)}{\text{odds}(HA^0 \mid women)} = \frac{.02/0.98}{0.08/0.92} = 0.235 \\
OR(HA) &= \frac{\text{odds}(HA^1)}{\text{odds}(HA^0)} = \frac{0.11/0.89}{0.29/0.71} = 0.303
\end{aligned}$$

Therefore the marginal OR is larger than that in either stratum!

The main take-away is that coefficients in general non-linear models are conditional and do not correspond to marginal (entire-population) effects (even if the model is correct, which is probably unlikely). This subtlety is often missed.

For example, suppose we did a study and obtained the above data. If we fit a (misspecified) logistic regression assuming the ORs were constant for men and women, we would report an OR of ∼0.24, but this overstates the OR we'd see if we gave entire population treatment versus not (∼0.3). This problem can be exacerbated the more covariates there are, or of course if the model is misspecified.

However, this problem does not arise in a (correctly specified) linear model, e.g.,

$$\mathbb{E}(Y \mid X, A) = \beta_0 + \beta_1 A + \beta_2^{\mathrm{T}} X$$

If the model is correct, then

$$\beta_1 = \mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)$$

so the coefficient is a conditional effect.

However, under the linear model assumption and $A \perp\!\!\!\perp Y^a \mid X$

$$\mathbb{E}(Y^1 - Y^0) = \mathbb{E}\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\} = \beta_1$$

so the parameter is also a marginal effect.

We have seen that going after coefficients in nonlinear regression models can be sub-optimal in experiments. Namely, we have to assume the model is correct (a huge assumption) and, even if the model is correct, the coefficient in that case will be a conditional effect which does not correspond to a well-defined effect in the whole population.

In what follows we will discuss how to deal with the second issue, and then after that the first issue.

Bibliography

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

D. D. Boos and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods*. New York: Springer, 2013.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

D. A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, pages 237–249, 2008.

S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, pages 29–46, 1999.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B*, 79(4):1229–1245, 2017.

E. H. Kennedy, S. Balakrishnan, and M. G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics (to appear)*, 2019a.

E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019b.

J. Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.